Frequently Asked Questions

Experimental Dataset for Supertract-Based Census Tract HPIs

Version: June 2022

DATASETS:

- https://www.fhfa.gov/PolicyProgramsResearch/Research/PaperDocuments/wp2101-data-tract-dta.zip (zipped, STATA)
- https://www.fhfa.gov/PolicyProgramsResearch/Research/PaperDocuments/wp2101-data-tract-csv.zip (zipped, comma delimited)

FHFA WORKING PAPER:
http://www.fhfa.gov/papers/wp2101.aspx

Q1: What is measured by the House Price Index (HPI)?

A: The HPI provides a means to measure annual appreciation in single-family house prices over certain periods. It also provides housing and real estate economists with an analytical tool that is useful for estimating changes in the rates of mortgage defaults, prepayments and housing affordability in Census tracts. The HPI is a measure designed to capture same-unit changes, which is different from existing measures, like medians and averages, that cannot distinguish between changes in housing stock and house prices. We are able to focus directly on price changes because our methodology compares the value of the same house at two points in time.

In our dataset, we construct annual HPIs for Census tracts using a new "supertract approach" which we describe in the paper. This involves grouping adjacent tracts together until a minimum number of transaction threshold is met (see more in Q9). This approach is meant to minimize aggregation bias subject to this transaction count minimum constraint.

Q2: Are these official FHFA indices?

A: No, they should be considered as experimental or developmental indices. These tract-level HPIs have been produced as part of Working Paper 21-01 (http://www.fhfa.gov/papers/wp2101.aspx) to stimulate discussion and comment. Please cite the working paper when using the data so we can more easily follow how they are being used and make improvements if they are needed.

Q3: Are the tract HPIs made in a way that differs from the FHFA HPIs?

A: Yes. While the indices use the same fundamental "repeat-sales" or "repeat-transactions" methodology, the precise method is slightly different with statistical

methods that determine when data should be pooled together into combined sample areas. Working Paper 21-01 has a detailed description on the data filters and methodology.

Q4: What data are used to make the indices?

A: The sample dataset begins with purchase-money mortgage originations purchased or securitized by Fannie Mae or Freddie Mac. Additional transactions are appended from Federal Housing Administration mortgages and county recorder purchase information from CoreLogic. We use a statistical method to measure the annual price change in repeated sales on the same property with data that extend back to the mid-1970s.

Q5: How could I adjust the nominal house price appreciation rates to real terms?

A: For inflation adjustments, you might consider the Consumer Price Index produced by the Bureau of Labor Statistics (http://data.bls.gov/cgi-bin/srgate). The most common series used for these adjustments are the "all items" (CUUR0000SA0) and "all items less shelter" (CUUR0000SA0L2).

Q6: Are the indices smoothed or adjusted for seasonality?

A: No, we have not adjusted the indices in any way beyond meeting certain sample requirements for the number of paired transactions. We offer raw data series so that researchers and analysts can make adjustments as they deem appropriate.

Q7: What transaction date is used to estimate the indices?

A: The transaction date is used, as opposed to the contract signing date or the loan acquisition date, which are sometimes used in other studies.

Q8: How might one connect demographic data with the Census tract HPIs?

A: The tracts are based on 2010 definitions, so any series from the Census with a GEOID from this time period can be used. Also, see nhgis.org for Census data for different time periods harmonized to 2010 definitions.

Q9: What is a "supertract"?

A: A supertract is a grouping of census tracts in a particular year for a given CBSA. It exists to group together adjacent tracts for a brief time period (in our case, a single year) to ensure a minimum number of transactions are available to estimate a price index. Supertracts can be as small as a single tract or as large as an entire city in cases of low transaction counts. They exist for one period and then are re-formed if needed in other years. There is a supertract ID that identifies tract

groupings in a particular year. Once they are used to estimate price indices for that grouping of tracts for that particular year, they are no longer used. In different years, a supertract may exist that consists of the same or different groups of tracts. See paper text and Q15 for more details.

Q10: How do these tract-level estimates compare to Bogin, Doerner, and Larson (2019)?

A: This new index database is highly correlated with the Bogin, Doerner, and Larson (*Real Estate Economics,* 2019) "BDL" index database, but with higher index counts. It is also based on different source data (this includes purchases from Fannie Mae, Freddie Mac, Federal Housing Administration, and county recorder offices; BDL's source data is Fannie Mae and Freddie Mac purchase and refinance mortgages). See Appendix table A1 in the paper, reproduced below.

Decade	Tracts (this paper)	Tracts (BDL Dataset)	Tracts (Intersection)	Annual Tract Appreciation Correlation
1990-1999	63,122	23,664	17,494	0.43
2000-2009	63,122	46,003	33,517	0.75
2010-2019	63,122	53,181	37,413	0.51

Q11: How can I view these data?

A: The data are available as .CSV and .DTA (Stata) files. However, be careful when opening the .CSV version in Excel or other applications that may have row limits which truncate observations. A proper import should have 2,083,026 rows (not counting the variable names in the .CSV file). Note that the .DTA version has metadata stored in the file that can be accessed by typing commands like describe or notes list in Stata.

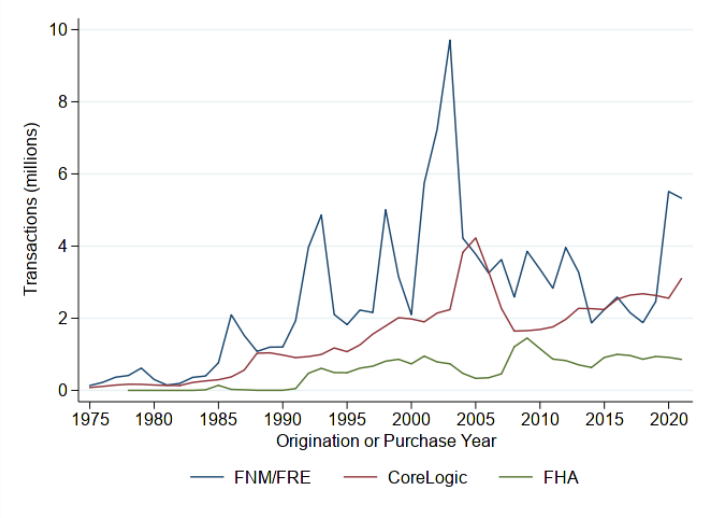
Q12: What filters are in the data?

A: We drop all unit-level transaction pairs with appreciation rates greater than/less than about 35% per year (log-difference of 0.3) between transactions, as is standard. This reduces the likelihood a quality change in the housing unit itself has occurred (i.e. major damage or major upgrade). To estimate an index for a city, there must be at least 40 transactions in every year between 1989 and 2021.

Q13: How do different data sources change over the sample period?

A: There are three sources of data for this study, including mortgages purchased or securitized by Fannie Mae and Freddie Mac (blue), CoreLogic county recorder files (red), and FHA mortgages (green). "Transactions" refer to mortgage-financed purchases or refinances (FNM/FRE/FHA), or transactions that result in

the purchase of a home (CoreLogic). The year is the date of loan origination (FNM/FRE/FHA) or purchase (CoreLogic). After duplicates are removed and filters are applied, we arrive at our database used to construct the indices.



Q14: Why are there some very large/small values in the appreciation rate database?

A: The indices are estimated at a very granular level, so there will be cases where statistical noise is high, which tends to happen more often in the earlier periods of data series. These high appreciation rates are typically offsetting in repeat-sales indices (see Clapp, Giaccotto, and Tirtiroglu (1991) in *Real Estate Economics*), though this has not been established for our current index rendition which has a slightly different estimation procedure that uses changing geographies to estimate indices.

Q15: How should I use the "supertract" variable?

A: This variable is defined as a CBSA-by-year identifier. Accordingly, supertract IDs with the same value do not represent the same supertract in different CBSAs or different years (even within the same CBSA). The purpose of the inclusion of this identifier is to allow researchers to understand and assess the pooling of tracts when the methodology in the paper is implemented (for instance, by mapping as in Figure 3.A or counts over time as in Figure A.2).

Q16: What revisions can we expect for this database?

A: This database is currently under development, with the working paper corresponding to the June 2022 version of the database. This version includes annual tract-level indices from 1989 through 2021. We may continue to revise the dataset with changes to supertract cutoffs, data filters, or other changes. The most current version of the database includes 63,122 tracts in 581 CBSAs.

Q17: Are there variable descriptions for this dataset?

A: Yes. See below for a description of the variables in the dataset. These descriptions are also provided in the .DTA (Stata) version of the dataset using the "notes" command.

Variable Name	Description
year	Year of index
cbsa	Core-Based Statistical Areas come from the Office of Management and Budget (OMB)'s March 2020 definitions.
tract	Tract is an 11-digit identifier (2010 definitions): State (2 digit) + County (3 digit) + Census Tract (6 digit).
hpi	HPI is the level index, which is normalized to 100 in 1989.
appr	Appreciation as percent change in level indices. For a given tract at time t, $appr_t = (hpi_t - hpi_{t-1})/hpi_{t-1}$
appr_ln	Appreciation as difference in logged indices. For a given tract at time t, $appr\_ln_t = \ln(hpi_t) - \ln(hpi_t)$
suptract	Supertract grouping number associated with the tract (varies by year and cbsa). All tracts with same *suptract* value are in the same supertract.